Collocational behaviour in non content words

Ton van der Wouden
NWO/Universities of Groningen, Leiden and Utrecht
UiL-OTS, Trans 10, 3512 JK Utrecht, The Netherlands
vdwouden@let.rug.nl

Abstract

The paper aims at serving two goals. Firstly, attention will be drawn to a class of collocations that, to our knowledge at least, has gone unnoticed so far. Secondly, it will be argued that brute force techniques as they are employed by certain automated collocation search strategies using a quantitative technique plus a mechanism to filter out unwanted combinations is a guaranteed way to miss this class of collocations.

1 Introduction

Under the most general interpretation of the notion COLLOCATION, any two lexical elements occurring more (or less, cf. below) often in each other's neighborhood than chance predicts should be considered having a collocational relationship. From this perspective, combinations such as *collect stamps* and *proud of* qualify as collocations, but the same holds for *have to* and *an apple* (Van der Wouden 1997, Part I).[1]

Cases such as the latter two show the weakness of a purely quantitative approach to collocation. Still, for lack of a better one that can be operationalized, quantitative definitions of collocation are often used, especially in the automatic retrieval of such fixed combinations.

Two popular strategies that have been employed to reduce the number of uninteresting (since transparent) combinations such as *an apple* and *have to* are part of speech filters and stop word lists.

- part of speech filters: only let through those syntactic structures that are likely to be 'phrases'.
- stop word lists: neglect certain words (usually high frequency function words) as parts of higher than chance bigrams and N-grams.

We will argue that these strategies of excluding certain high frequency elements from the set of potential collocations is not without its dangers. More particularly, we will demonstrate that, at least in Dutch, certain high frequency function words show all kinds of collocational effects. Part of speech filters and stop word lists in the usual sense will effectively filter out these effects.

2 Collocations

Let us start with a very general notion of collocation, e.g. along the following lines

(1) The term COLLOCATION refers to the idiosyncratic syntagmatic combination of lexical items and is independent of word class or syntactic structure (Fontenelle 1992: 222)

More often than not, the definition adopted in collocation research, be it theoretically oriented or computational, is far more restricted than this one. We give one example of such a definition from

[1] This paper elaborates on ideas presented in my (2000a). The research reported on here fits into the larger framework of the VNC-project "Partikelgebruik in Nederland en Vlaanderen", financed by NWO, the Dutch Organization for Scientific research, and her Belgian sister organization FWO.

(Hausmann 1989–91), in which both the restriction to certain syntactic structures and the lack of motivation thereof are particularly striking:

> “On appelera collocation la combinaison caractéristique de deux mots dans une des structures suivantes : a) substantif + adjectif (épithète) ; b) substantif + verbe ; c) verbe + substantif (objet) ; d) verbe + adverbe ; e) adjectif + adverbe ; f) substantif + (prép.) + substantif. La collocation se distingue de la combinaison libre (*the book is useful / das Buch ist nützlich / le livre est utile*) par la combinabilité restreinte (ou affinité) des mots combinés (*feuilleter un livre* vs. *acheter un livre*).

One of the aims of this contribution is to show that lexical elements of almost any class may show collocational effects (cf. also my (1997)).

We will try to demonstrate this with the Dutch (as opposed to Belgian) subpart of the second release of the Dutch Spoken Corpus, which is currently under development.[2] The size of this Dutch subcorpus is currently just over a million words, but that is large enough for our main argument.

3 Quantitative implementions of collocation

Even if we skip discussion of the problems involved in terms such as IDIOSYNCRATIC (cf. Van der Wouden 1997, Part 1), definitions of collocation such as the one given in (1) are very hard to implement. In computational linguistics, a common way of approaching the linguistic concept of collocation is quantificational. The simplest way of course is to just look at frequently occurring combinations.

As an efficient means to investigate this corpus, we used the “Bigram Statistics Package” (BSD) developed by Ted Pedersen and Satanjeev Banerjee of the University of Minnesota, Duluth. This is a library of Perl routines allowing the researcher to efficiently extract bigrams from a corpus and apply various statistical metrics on these bigrams, among other things.[3]

The usefulness of the BSP tools can be demonstrated with two textbook examples of collocations, the fixed prepositions that come with the verb *luisteren* ‘listen’, to wit, *naar* and the one for the adjective *dol* ‘fond’, which is *op*.[4]

The program found two bigrams with the string *luister* (which occurs 51 times in the corpus) and one with *dol* (text frequency 12):[5]

(2)

bigram	*N*
dol op	5
luister naar	6
. luister	5
luister eens	5

It is clear that this is quite a good result, as *luister naar* and *dol op* were exactly the bigrams we were looking for in the first place. The unexpectedly high frequency of *. luister* indicates that *luister* is often the first word of an utterance (in this corpus). *Luister eens* is unexpected as well but after inspection of the examples this combination, and its relatively high frequency, can be explained. Essentially, we are dealing here with usage of *luister* ‘listen’ as an imperative (morphology doesn’t distinguish between first person singular and imperative in most Dutch verbs). The directive force of the imperative is mitigated (Vismans 1994) to the effect that the combination *luister eens* functions more as a suggestion than as a true command. Moreover, as the example in (3) shows, the combination *luister eens* need not be part of

[2] The aim of the Spoken Dutch Corpus project (abbreviated as CGN, from the Dutch name *Corpus Gesproken Nederlands*) is to build an annotated corpus of about one thousand hours of continuous speech, which amounts to 10 million words. The project started in June 1998, and runs for five years. It is a collaborative effort of several Dutch and Flemish universities (Oostdijk 2000, Hoekstra et al. 2001).

The corpus is intended as a major resource both for linguistic research and for language and speech technology. To serve this dual purpose, it contains materials recorded in a variety of communicative settings: spontaneous face-to-face and telephone dialogues, interviews, discussions, debates, lectures, news broadcasts and book passages read aloud. Two-thirds of the material is collected in the Netherlands, one third in the Dutch speaking part of Belgium. Upon completion, the corpus will be the largest and most diverse database of spoken Dutch collected so far. Cf. the project’s website http://lands.let.kun.nl/cgn.

[3] The BSD programs are free software under the terms of the GNU General Public License; the code can be found at the internet address http://www.d.umn.edu/∼tpederse/code.html.

[4] For reasons of presentation, we restrict ourselves to adjacent two-word collocations, i.e. bigrams, until section 6.

[5] Arbitrarily, we only take into account bigrams that occur at least five times in the corpus. Note that the program interprets punctuation marks such as full stops as words too.

a larger syntactic unit. In this case, it appears to function as a kind of discourse marker in the sense of (Schiffrin 1987) and (Fraser 1999): the speaker expresses the wish to make an statement and does not want an intervention. That is, (s)he is taking the floor and tries to keep it.

(3) *maar Dick luister eens als wij*
but Dick listen PART if we
nou een vakantie d'raan vastplakken
PART a holiday it-to glue
daar in Nieuw-Zeeland
there in New Sealand
'but Dick, listen please. Suppose we combine it with a holiday in New Sealand'

Just looking for frequent combinations is not always the most useful way to look for possible collocations. More sophisticated techniques compare the frequency of the combination with the frequency expected, given the frequencies of the parts constituting the combination (Manning and Schütze, 1999, Ch. 5). For this purpose, the BSP tools come with a number of statistical tests. We will not go into the statistical background of any of these tests, we just give the outcome in (4):[6]

(4)

bigram	N	χ^2	ll	mi
luister naar	6	1272.3	56.5	8.0
luister eens	6	1235.0	47.9	7.7
. luister	5	33.3	20.3	2.9
dol op	5	805.1	58.9	7.1

We see that in all tests, *luister naar* has the highest score and *. luister* the lowest. Moreover, the scores of *luister naar, luister eens* and *dol op* are all in the same order of magnitude, while the value for *. luister* is much lower consistently. This is exactly what we want, of course, as *luister naar* and *dol op* fit more into our pretheoretic notion of collocation than any combination of a word and a punctuation sign.

4 Collocational behaviour of focus particles

As was already indicated above, we believe that collocation is not restricted to content words such as nouns, verbs and adjectives. We will now show that the techniques just sketched yield collocational effects in other word classes as well.

Consider for example the Dutch restrictive focus particle *alleen* 'only', which occurs no less than 1120 times in the corpus. 75 bigrams with *alleen* pass the threshold of five or more occurrences. The bigrams that score highest in some of the statistical tests are listed in (5).

(5)

bigram	N	χ^2	ll	mi
alleen maar	327	9845	1782	5.0
niet alleen	201	3548	864	4.3
alleen nog	32	156	68	2.7
helemaal alleen	11	121	36	3.7

These results are not without interest. The combinations ranked highest, viz. *alleen maar* and *niet alleen*, were already mentioned in 1898 in the entry *alleen* of the historical dictionary *Woordenboek der Nederlandsche Taal*. In *alleen maar*, *maar*'s main function seems to be a form of rhetorical strengthening (6), whereas *niet alleen* has two important uses: in cases as examplified in (7), the meaning is more or less compositional, whereas in cases such as (8) it is part of a larger coordination construction *niet alleen . . . maar ook* 'not just . . . but . . . as well'.

(6) *'t gaat alleen maar over treinreizen*
it goes only but about train trips
'it is about train trips only'

(7) *ik ga niet alleen les geven*
I go not only lesson give
'I'll not just teach'

(8) *niet alleen vanuit commercieel*
not only from commercial
oogpunt maar ook wetenschappelijk
viewpoint but also scientifically
gezien
seen
'both from a commercial and a scientific point of view'

From a linguistic point of view, it can be argued that the usages examplified in (6–8) should indeed be classified as collocations proper on the basis of

[6] χ^2 is the BSP implementation of the familiar χ^2 test, **ll** is log-likelihood, **mi** is mutual information. For reasons of lack of space we omit some of the less significant decimals.

criteria such as the ones quoted in (Manning and Schütze, 1999, 184):[7]

- non-compositionality: the particle *maar* (on which (Foolen 1993)) can never be used to rhetorically strengthen an adverb, except for the case of *alleen* (and for some speakers, its near-synonym *enkel*);

- non-substitutability: for Dutch (as opposed to Belgian) speakers, *alleen* in (6) and in (8) cannot be substituted by any of its (near) synonyms such as *enkel* or *slechts* (Van der Wouden 2000b);

- non-modifiability: usually, both *niet* 'not' and *alleen* 'only' can be modified by adverbs of degree such as *vrijwel* 'almost' (Klein 1997). Although this *vrijwel* can be added felicitously to (6) (*'t gaat vrijwel alleen maar over treinreizen* 'it is almost only about train trips'), it cannot in the cases of (7) (**ik ga vrijwel niet alleen les geven*, **ik ga niet vrijwel alleen les geven*) and (8) (**vrijwel niet alleen vanuit* ...).

The last two combinations in the table in (5) are perhaps not as collocational as the first three. *Alleen nog* is a combination of *alleen* with the temporal adverb *nog*, which is more or less comparable to English *still* and *yet*. The combination *alleen nog* is typically used to express that there is one thing left, with the suggestion that there used to be more (Van Baar 1997):

(9) *we moeten alleen nog het probleem*
we must only yet the problem
van het geluid oplossen
of the sound solve
'The sound is the only problem left to be solved'

(10) *ik wil alleen nog maar dood*
I want only PART PART dead
'there's just one thing left that I want: to die'

Finally, *helemaal alleen* is a combination of a completely different kind, as *alleen* functions as a predicate here (English *alone*), for which *helemaal* 'totally' is the protypical (collocational?) adverb of degree:

(11) *hij was helemaal alleen naar het*
he was totally alone to the
grote gebouw gelopen
big building walked
'he had walked to the big building all by himself'

5 Collocational behaviour of modal particles

Comparable effects may be found in the so-called 'modal particles', an infamous class of adverb-like elements which are hard to define, have a hard to describe, often context-sensitive contribution to the meaning of the sentence or utterance, but are quite common in e.g. the mainland Germanic languages (cf., e.g., (Abraham 1991)).

We will demonstrate this with the particle *eens*, the cognate of English *once* we already met in section 3. The original meaning of *eens* is comparable to that of *once*, viz., an existential quantifier over time. This function, however, has been largely taken over by other lexical elements (Zwarts 2000); in most cases, *eens* functions as a particle nowadays.

The string *eens* occurs in the corpus 1633 times.[8] The table in (12) lists the six combinations with *eens* occurring most, plus the scores for three statistical tests.[9]

(12)

bigram	N	χ^2	ll	mi
wel eens	426	13789.1	2336.0	5.09
eens een	240	1625.3	645.2	3.09
nog eens	207	4696.4	964.2	4.61
niet eens	148	894.1	364.4	2.97

The meaning and the usage possibilities of the first and most frequent combination, *wel eens*, in

[7] According to the definition in (1), all combinations occurring either more or less often than expected qualify as collocational.

[8] When used as a modal particle, *eens* is almost always unstressed (cf. (Van der Wouden et al., 1998) and the literature given there). Unofficial spellings such as *'ns* are meant to express this unstressed use. One finds as many cases of *'ns* as of *eens* in the corpus, but the orthographic transcription of the CGN is inconsistent in the choice of *'ns* vs. *eens*. As there is no other usage of *'ns*, we could safely remove this inconsistency by changing all cases of *'ns* into *eens*.

[9] Time, space, nor our competence permit comparison of these scores with the ones for the text book collocations *luister naar* and *dol op* in (4) in order to decide which of the statistical tests is best, or which of the combinations are most collocational in a statistical sense.

which *eens* is preceded by another modal particle *wel*, are very hard to describe. *Eens*'s contribution is often slightly temporal ('once, once upon a time, sometime, someday'), *wel* is somewhat concessive ('I admit . . . '); in questions it is often best translated as *ever*. In quite a number of cases it is impossible to leave out one of the parts without getting a result that is either ungrammatical or has a completely different meaning:

(13) *dat is wel eens lastig*
that is PART PART hard
'that may be hard sometimes'

(14) **dat is eens lastig*

(15) *dat is wel lastig*
that is PART hard
'I admit that it is hard'

This indicates that the meaning of *wel eens* is quite idiomatic. Another argument in favour of considering *wel eens* to be a collocation is the fact that various sources, including the *Woordenboek der Nederlandsche Taal* and the influential *Schrijfwijzer* (Renkema 1989), prescribe that *wel eens* in this sense be written as *weleens*, i.e., as a word. In the orthographic transcription of the CGN corpus, this happened only 12 times. (16–17) are two examples:

(16) *ik eet ook weleens wat op 't*
I eat also PART something at the
station maar bijna nooit.
station but almost never
'I do sometimes have a snack at the station but very rarely'

(17) *is het u weleens opgevallen hoe*
is it you PART noticed how
gespannen u wordt?
tensed you become
'did you ever notice how tensed you become?'

The next combination in (12) is *eens een*, in which *eens* is followed by the indefinite article *een* 'a'. This combination does not look like a collocation – but cf. (22) below.

Next comes *nog eens*, in which *eens* is preceded by temporal *nog*. Depending on tense and mood of the sentence, this combination may translated as *once again* or *ever*, among other things:

(18) *misschien worden ze nog eens*
perhaps become they PART PART
wakker
awake
'perhaps they'll wake up some time'

(19) *schenk me nog eens een borrel in*
pour me PART PART a drink in
'pour me out a drink once again'

It is unclear whether to classify this frequent combination as collocational on linguistic grounds.

The combination *niet eens* definitely qualifies as a collocation: the meaning is completely non-compositional, the original semantics of *eens* is lost. The combination functions as a negative focus particle, comparable to English *not even*.[10] Cf. the example in (20): in the normal case, a verb such as *weten* 'know' cannot be combined with an existential quantifier over time, as one cannot know something once.

(20) *ik weet het niet eens*
I know it not PART
'I don't even know it'

6 More complex collocational behaviour: beyond bigrams

Collocational behaviour with adverbial elements is not restricted to bigrams: on the one hand, collocational relations may exist between more than two words, and lexical elements having a collocational relationship need not be adjacent (a textbook example of a long distance collocation is *collect stamps*: the collocates may be arbitrarily far apart, as in *the only thing I ever collected in my entire life, apart from . . . , are stamps*). The BSP tools offer an option to extend the "window" of possible collocants from the standard value 2 (which means that only the words immediately to the left and to the right of the keyword are taken

[10] It is surprising that *niet eens* is rarely written as one word. Here is one rare example: *De vrouw wier naam hij nieteens wist. . .* 'The woman whose name he even didn't know (Albert Helman, *Het vergeten gezicht*. Rotterdam: Nijgh & Van Ditmar, 1939, p. 18.).

into account). The table in (21) lists some of the most frequent collocants of *wel eens* (written in any form) given a window of 10, that is maximally five words to the left or to the right of the keyword *wel eens*.

(21)

bigram	N	χ^2	ll	mi
. wel eens	378	0.25	0.25	0.04
wel eens .	334	1.25	1.27	-0.08
ik wel eens	192	203.7	141.0	1.41
wel eens een	123	61.7	48.1	0.99
ook wel eens	118	202.3	125.0	1.77
nog wel eens	76	219.0	115.6	2.22

The most frequent combinations, with the full stop, are completely uninteresting of course – a fact that seems to be reflected in the scores of the statistical tests. The next one, *ik wel eens*, is relatively uninteresting as well, *ik* being the first person pronoun nominative singular, i.e., the counterpart of *I*. *Wel eens een*, with the indefinite pronoun *een*, looks terribly boring again. However, 29 cases, or one fourth of the total, are instances of the combination *wel eens een keer* (24 cases) or its diminutive variant *wel eens een keertje* (5 cases). The expression *een keer*(*tje*) seems to function as a modal particle comparable to *eens* here.[11] An example is given below:

(22) *ik wil wel eens een keer een*
I want PART PART PART a
stuk van Dennis Potter zien
piece of Dennis Potter see
‘I’d like to see a Dennis Potter play one time’

The other two combinations involve indisputable particle clusters: with *ook* ‘also’ and with *nog* we met earlier. Examples are in (23–24):

(23) *dat zou ik ook wel eens*
that would I PART PART PART
willen ja
want yes
‘yes, I would like that as well’

(24) *niemand vroeg aan mij of ik nog*
no one asked to me if I
wel eens aan hem dacht
PART PART on him thought
‘no one asked me if I ever thought of him still’

It is known from the linguistic literature that modal particles cluster easily and frequently (cf., e.g., De Vriendt et al.), but the table in (21) shows that this tendency can be found back in the statistics as well – even in such a small corpus.

Next, we turn to *nog eens*.

(25)

bigram	N	χ^2	ll	mi
nog eens .	131	0.6	0.7	-0.09
. nog eens	118	4.2	4.4	-0.26
nog eens een	75	106.4	69.1	1.61
ik nog eens	62	44.1	32.7	1.17
ook nog eens	42	63.1	40.1	1.66
nog eens keer	41	1288.9	209.7	5.06

The most interesting combinations, both from a linguistic point of view and in light of the statistics, are again the ones with particles: *ook nog eens* on the one hand, and *nog eens een* and *nog eens keer*, which relate to *nog eens een keer*(*tje*):

(26) *dan wil ik ook nog eens*
PART want I PART PART PART
een krant lezen
a newspaper read
‘and then I want to read a newspaper too’

(27) *ik zal dat nog eens een keer*
I will that PART PART PART
nalezen
read through
‘I’ll read through that some time’

Finally, we turn to *niet eens*:

(28)

bigram	N	χ^2	ll	mi
niet eens .	170	67.2	56.2	0.84
. niet eens	139	18.5	16.7	0.49
ik niet eens	49	46.6	32.9	1.34
niet eens dat	39	8.5	7.2	0.66
dat niet eens	38	6.1	5.3	0.57
nog niet eens	35	215.4	85.5	3.0

The last one, with the particle *nog*, is the most interesting combination again; *dat* is a subordinating complementizer or a pronoun.

(29) *ik wist niet eens dat die bestond*
I knew not even that that existed
‘I even didn’t know that it existed’

[11] The χ^2 values for *wel eens keertje* and *wel eens keer* are 274.2 and 140.9, respectively, the loglikelihood scores are 30.7 and 56.6, and the mutual information values are 5.83 and 2.95.

(30) *dat wil je niet eens weten*
that want you not even know
‘you don’t even want to know that’

So far, we looked at the most frequent combinations around particle clusters. If, however, we look at the combinations ranked highest by the statistical tests we get interesting results as well – and a lot of noise. The table in (31) gives the combinations with *niet eens* scoring highest in the χ^2 test – but note that all combinations, except for *niet eens .*, score high on the other tests as well.

(31)

bigram	N	χ^2	ll	mi
nog niet eens	35	215.4	85.5	3.0
niet eens weten	8	199.0	37.3	4.75
niet eens meer	19	176.5	57.5	3.48
weet niet eens	16	134.1	45.9	3.36
weten niet eens	5	71.4	18.5	4.02
niet eens .	170	67.2	56.2	0.84

The combination scoring highest in this test, *nog niet eens*, was discussed already. No less than three combinations in this list of six involve forms of the verb *weten* ‘know’, so that certainly qualifies as a collocation (examples were already given above, e.g. in (29–30). *Meer* is again a particle-like element, a negative polarity item comparable to English *anymore*:

(32) *dat doen we niet eens meer*
that do we not even anymore
‘we don’t even do that anymore’

Space doesn’t permit us to give data about the collocational preference of the various particle combinations for modal auxiliaries (Van der Wouden, 2000a).

7 Against phrase filters and stop words

It is easy to automatically and efficiently search text corpora of considerable size for bigrams with occurrences higher than chance. However, automatically looking for high frequency bigrams yields more junk (e.g. determiner noun combinations) than collocations in the intuitive, linguistic sense. Therefore, mechanisms have been proposed to separate the interesting bigrams from the uninteresting ones.

A very simple and succesful heuristic is to pass the candidate phrases through a part of speech filter which only lets through those patterns that are likely to be ‘phrases’. (Ross and Tukey, 1975; Justeson and Katz, 1995), (Manning and Schütze, 1999, 153–5). For example, (Justeson and Katz, 1995) proposed the patterns (for English) given in (33):

(33)

Phrase filter proposed by
(Justeson and Katz, 1995)
(after (Manning and Schütze, 1999, 154))

Tag Pattern	Example
A N	linear function
N N	regression coefficients
A A N	Gaussian random variable
A N N	cumulative distribution function
N A N	mean squared error
N N N	class probability function
N P N	degrees of freedom

As (Manning and Schütze, 1999, 155) state, “The results are surprisingly good”: the bigrams with a relatively high frequency that passed the filter comply with their (and our) intuitive sense of collocation. However, and less surprisingly, the results are not that good if we look at this heuristic from our perspective, wishing to automatically find collocational behavior of particles and of function words in general. For that purpose, the filter turns out to be far too restrictive, i.e., completely useless. All non-standard types of collocations we have been discussing so far will be blocked, as they do not fit into the phrasal patterns contained in the phrase filter.

Addition of extra phrasal patterns to account for the types of collocations discussed here would bring back the enormous amounts of “junk bigrams” that do not correspond to our intuitions about what should count as a collocation. This pessimism is inspired by the fact that particles would probably be tagged ‘Adv’, which would mean that we would be looking for ‘Adv Adv (Adv (Adv))’ patterns. But then all kinds of uninteresting adverbial bigrams such as *tomorrow again*, *early tonight*, *actually lower* and *much more likely* would pop up as well. It would again be very hard to select the interesting collocations from the uninteresting ones.

Another filtering method proposed (Smadja and McKeown, 1990) and used is a list of so-called “stop words”, usually high frequency functions words, which are to be neglected as parts of higher than chance bigrams and N-grams. Again, this method has proven to be successful for "classical" collocations (Manning and Schütze, 1999) and again, this method is useless for the kinds of collocations we've been discussing her, as that is exactly what most of the particles and other elements involved are (among other things): high frequency functions words.

References

Werner Abraham. 1991. Modal particle research: The state of the art. *Multilingua*, 10:9–15.

Tim van Baar. 1997. *Phasal Polarity*. Ph.D. thesis, Universiteit van Amsterdam.

Thierry Fontenelle. 1992. Collocation acquisition from a corpus or from a dictionary: a comparison. In Hannu Tommola, Krista Varantola, Tarja Salmi-Tolonen, and Jürgen Schopp, editors, *EURALEX '92 Proceedings I–II. Papers submitted to the 5th EURALEX international congress on lexicography in Tampere, Finland*, pages 221–228. Department of translation studies, University of Tampere.

Ad Foolen. 1993. *De betekenis van partikels. Een dokumentatie van de stand van het onderzoek met bijzondere aandacht voor* maar. Ph.D. thesis, Nijmegen.

Bruce Fraser. 1999. What are discourse markers? *Journal of Pragmatics*, 31:931–952.

Franz Josef Hausmann. 1989-1991. Collocations. In Franz Josef Hausmann et al., editors, *Wörterbuecher: ein internationales Handbuch zur Lexikographie / Dictionaries: an international encyclopedia of lexicography / Dictionnaires: encyclopedie internationale de lexicographie*, pages I: 1010–19. De Gruyter, Berlin [etc.]. (Handbücher zur Sprach- und Kommunikationswissenschaft; Bd. 5).

Heleen Hoekstra, Michael Moortgat, Ineke Schuurman, and Ton van der Wouden. 2000. Syntactic annotation for the spoken dutch corpus project (cgn). paper delivered at Computational Linguistics in the Netherlands, Tilburg, November 2000, accepted for publication in the Proceedings.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for indentification in text. *Natural Language Engineering*, 1:9–27.

Henny Klein. 1997. *Adverbs of degree in Dutch*. Ph.D. thesis, Groningen.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. The MIT Press, Cambridge, Mass.

Nelleke Oostdijk. 2000. Building a corpus of spoken Dutch. In Paola Monachesi, editor, *Computational Linguistics in the Netherlands 1999. Selected Papers from the Tenth CLIN Meeting*, pages 147–157. Utrecht University, Utrecht Institute of Linguistics OTS, Utrecht.

Jan Renkema. 1989. *Schrijfwijzer*. SDU uitgeverij, 's-Gravenhage. Volledig herz. ed. (Oorspr. uitg.: 1979).

Ian C. Ross and John W. Tukey. 1975. Introduction to these volumes. In John W. Tukey, editor, *Index to Statistics and Probability*, pages iv–x. R & D Press, Los Altos.

Deborah Schiffrin. 1987. *Discourse markers*. Cambridge University Press, Cambridge.

Frank A. Smadja and Kathleen R. McKeown. 1990. Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 252–259.

Roel Vismans. 1994. *Modal particles in Dutch directives: a study in functional grammar*. Ph.D. thesis, Vrije Universiteit Amsterdam.

Sera de Vriendt, Willy Vandeweghe, and Piet Van de Craen. 1991. Combinatorial aspects of modal particles in Dutch. *Multilingua*, 10:43–59.

Matthias de Vries and Lammert A. te Winkel et al., editors. 1864–1998. *Woordenboek der Nederlandsche taal*. Martinus Nijhoff [etc.], 's-Gravenhage [etc.].

Ton van der Wouden, Frans Zwarts, Inge Callebaut, and Piet Van de Craen. 1998. Once upon a time in Dutch. Ms. Leiden/Groningen/Brussel, 1998, URL: http://www.let.rug.nl/˜vdwouden.

Ton van der Wouden. 1997. *Negative Contexts. Collocation, polarity, and multiple negation*. Routledge, London and New York.

Ton van der Wouden. 2000a. Collocational behaviour in the modal realm. Paper presented in the workshop on Collocation, DGFS-meeting, March 2000, Marburg.

Ton van der Wouden. 2000b. Prototypicality vs. variation: Restrictive focus particles in Dutch. Paper delivered at Discourse Particles, Modal And Focal Particles And All That Stuff . . . , An international conference on Particles, Brussels, 8–9 December 2000.

Frans Zwarts. 2000. Dutch as a Davidsonian language. Paper delivered at Discourse Particles, Modal And Focal Particles And All That Stuff . . . , An international conference on Particles, Brussels, 8–9 December 2000.